

Master 2 Biologie-Informatique/ Bioinformatique



Classification pour l'étude des séquences répétées centromériques chez les primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Muséum national d'Histoire naturelle, 43 rue Cuvier 75005 PARIS



Remerciements

Je tiens tout d'abord à remercier énormément Loïc Ponger, responsable de mon stage, pour son encadrement, ses conseils, ses relectures et son aide.

Je tiens également à remercier Christophe Escudé pour tous ses conseils et pour la relecture de mon rapport, ainsi que l'équipe ARChE.

Je souhaite aussi remercier Evelyne Duvernois-Berthet, pour ses conseils tant au niveau professionnel que personnel, pour les discussions et pour les bonbons.

Je remercie aussi chaleureusement tout le laboratoire pour son accueil cordial.

Je remercie le journal club pour ces petites séances d'anglais sympathiques.

Je souhaite également remercier ici Catherine Etchebest et Jean-Christophe Gelly, et l'ensemble de l'équipe pédagogique, pour cette année de master 2.

Table des matières

Remerciements	1
1 Introduction	1
1.1 Le centromère	1
1.2 L'ADN α -satellite	1
1.3 Le sujet de stage	3
2 Matériel et méthode	4
2.1 Les espèces étudiées	4
2.2 Méthode de classification	4
2.2.1 Principe	4
2.2.2 Répartition itérative	5
2.2.3 Double-validation d'un sous-groupe	5
2.3 Analyse des séquences	6
3 Résultats	6
3.1 Caractérisation intraspécifique des familles	6
3.1.1 Identification des familles	6
3.1.2 Motifs potentiellement fonctionnels	11
3.1.3 Similarité entre familles	12
3.2 Comparaison inter-espèce	13
3.2.1 Répartition des super-familles	14
3.2.2 Un groupe C2 hétérogène	15
3.2.3 Origine du motif pK β	16
4 Discussion	17
5 Conclusion	19

1 Introduction

1.1 Le centromère

Le centromère est une structure chromatinienne essentielle au bon déroulement de la division cellulaire chez les eucaryotes. Il permet l'attachement du fuseau mitotique et la ségrégation des chromosomes [1]. Le kinétochore est un assemblage de protéines situé au niveau du centromère. Il va permettre l'attachement des microtubules aux chromosomes [2]. La chromatine centromérique est caractérisée par la présence de la protéine CENP-A, un variant de l'histone H3, très conservé au cours de l'évolution et qui détermine la position du kinétochore par un mécanisme épigénétique encore mal connu [3].

Bien que la fonction du centromère et des protéines associées soit relativement bien conservée, les séquences d'ADN associées varient d'un taxon à l'autre. Lorsque cet ADN est comparé entre plusieurs espèces, les séquences sont très différentes l'une de l'autre. Mais une structure commune se distingue : des motifs d'ADN répétés en tandem appelés ADN satellite [4]. Ces séquences peuvent représenter environ 5% du génome et la taille des unités de répétitions peut varier entre 7pb et 3,2kb [5] chez les mammifères.

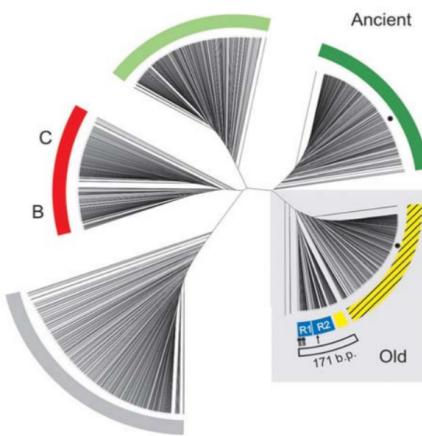


FIGURE 1 – Arbre phylogénétique des séquences α -satellites du bras p du chromosome X chez l'Homme : L'arbre réunit 1431 monomères présents sur le contig nt011630 [6].

1.2 L'ADN α -satellite

Chez les Primates, l'ADN satellite le plus abondant est connu sous le nom d' α -satellite. Ces séquences sont riches en AT et un monomère fait 171 pb de long environ [7]. Ces séquences ont été mises en évidence pour la première fois chez *Chlorocebus aethiops* dans les années 1970

[8]. Des homologues ont été retrouvés chez d'autres espèces de primates [?]. Néanmoins, ces séquences ont été essentiellement étudiées chez l'Homme.

Les séquences ont un taux d'identité qui varie de 60 à 100% [9]. Les séquences les plus similaires peuvent être regroupées en familles (Figure 1). Ces familles résultent d'événements d'amplification. Des études chez l'Homme ont montré que les séquences d'une même famille se regroupent phylogénétiquement mais aussi spatialement le long d'un chromosome [6]. Les familles se disposent de façon symétrique autour du cœur du centromère en respectant un gradient d'âge. Ces observations ont permis de proposer un modèle évolutif avec des centromères en expansion. Les familles les plus récentes s'inséreraient au cœur du centromère, dans la partie active, repoussant les familles les plus anciennes jusqu'aux régions voisines, appelées péri-centromères. Les familles les plus anciennes ont une organisation monomérique, sous forme de longs segments composés de monomères de la même famille. Les familles récentes sont organisées en structure d'ordre supérieur ou *Higher Order Repeat* (HOR), où un groupe de monomères appartenant à des familles différentes est répété en bloc en tandem (Figure 2).

Le rôle des α -satellites est encore mal connu dans la fonction du centromère, mais des études chez l'Homme ont permis de mettre en évidence l'existence des motifs conservés : la boîte CENP-B et la boîte pJ α . Le motif CENP-B pourrait lier la protéine CENP-B, une protéine capable de reconnaître et de se fixer sur de l'ADN. Le motif pJ α serait capable de lier la protéine pJ α , qui très mal caractérisée. Un troisième motif conservé nommé pK β a récemment été mis en évidence dans l'équipe (travaux non publiés). Ces trois motifs sont exclusifs l'un de l'autre et se situent au même endroit sur une séquence.

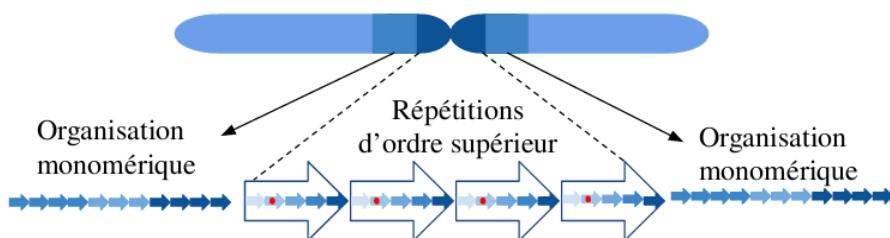


FIGURE 2 – Organisation spatiale des α -satellites : Le cœur du centromère (bleu foncé) est organisé en répétitions d'ordre supérieur. Le péricentromère (bleu clair) a une organisation monomérique. Les monomères sont représentés par une flèche bleue. Les flèches d'une même nuance de couleur appartiennent à la même famille. Les points rouges représentent les motifs CENP-B, pJ α ou pK β .

1.3 Le sujet de stage

L'objectif de ce stage est d'étudier les α -satellites à partir de données de séquençage haut débit afin de mieux comprendre la fonction et les mécanismes d'évolution de ces séquences.

Peu d'informations sur les α -satellites existent chez les espèces de primates. De plus, aucune comparaison inter-espèce n'a été réalisée, l'assemblage de ces séquences étant difficiles de par leur caractère répété. Jusqu'à présent, des études à grande échelle ont été appliquées chez l'Homme [9], utilisant du séquençage haut-débit, et chez le Gorille [10], utilisant du séquençage moyen-débit. L'équipe d'accueil de mon stage "ADN répété, Chromatine, Evolution" (ARChE), a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α -satellites chez *Cercopithecus solatus* [11] et *Cercopithecus pogonias* [12]. Ces deux clades ont été choisis car ils présentent beaucoup d'ADN satellite et de nombreux réarrangements chromosomiques, avec l'apparition de nombreux centromères. Une des étapes initiales à l'étude de ces séquences est de les classer en familles. La classification de ces séquences en utilisant des alignements et de la phylogénie n'est pas possible sur des jeux de données aussi grands. Or pour caractériser la diversité des α -satellites, il faut prendre en compte plusieurs milliers de séquences. De plus, elles ne permettent pas de faire des comparaisons entre espèces.

Pour remédier à ce problème, une méthode de classification automatisée des α -satellites a été développée dans l'équipe (Jornod, Master 2, 2016-2017). Elle permet de traiter des centaines de milliers de séquences, de les classer en familles sans avoir à déterminer initialement le nombre de familles. De plus, cette méthode est objective et peut être appliquée à plusieurs espèces, permettant ainsi une comparaison inter-espèce des familles α -satellites.

Afin d'évaluer cette méthode, je vais l'appliquer aux jeux de données de Cercopithèques déjà publiés par l'équipe. Dans un second temps, cette méthode est appliquée à deux autres primates dans le but de caractériser les familles d'espèces proches. Les résultats seront replacés dans le contexte phylogénétique de ces 4 espèces afin de comprendre un peu mieux l'histoire évolutive de ces séquences.

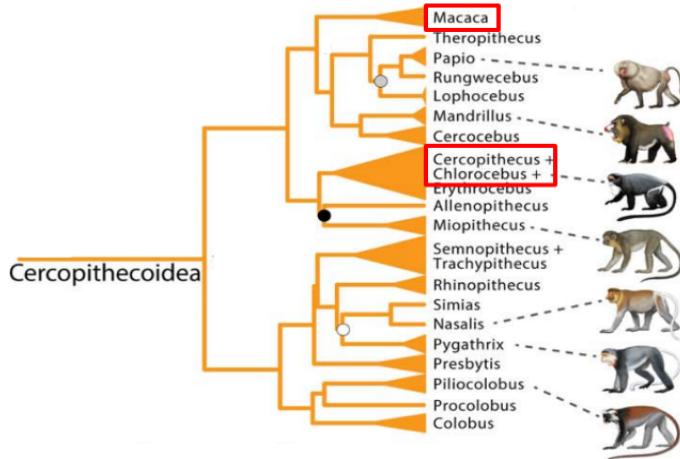


FIGURE 3 – Arbre phylogénétique des espèces analysées.[13]

2 Matériel et méthode

2.1 Les espèces étudiées

Les données des deux Cercopithèques, *C. solatus* et *C. pogonias*, sont issues d'un séquençage ciblé de monomères digérés par une enzyme de restriction (XmnI, HindIII) et séquencés en Ion Torrent [14].

Ensuite, une recherche d' α -satellites chez d'autres primates, issus du séquençage à haut débit, a été réalisée sur Genbank. Deux espèces proches des deux Cercopithèques, *Macaca fascicularis* et *Chlorocebus sabaeus*, sont choisies parmi 10 autres espèces. Elles ont été sélectionnées selon la qualité des séquences disponibles.

Les monomères d' α -satellites faisant environ 171pb, seuls les projets impliquant des reads relativement longs ont été considérés (Illumina en single-end, L454, ...). Tous les α -satellites sont alignés (même phase) sur les monomères de *C. solatus* et *C. pogonias* étudiés précédemment.

2.2 Méthode de classification

2.2.1 Principe

Cette méthode a été développée par Florence Jornod (stage M2, 2016-2017). Elle répartit des séquences α -satellites en familles selon la similarité calculée à partir de leur composition en 5-mers. La classification est hiérarchique dichotomique. Au départ, une table contenant les fréquences des 5-mers est calculée pour chaque monomère. Ensuite une boucle itérative est

exécutée pour séparer les séquences en deux groupes (voir 2.2.2). Cette étape est répétée tant que les nouveaux groupes formés sont divisibles (schéma de l'algorithme en annexe).

2.2.2 Répartition itérative

Une Analyse en Composante Principale (ACP) est effectuée sur la table des fréquences des 5-mers afin de réduire les dimensions du jeu de données et d'obtenir des variables indépendantes. Les distances euclidiennes sont calculées entre toutes les paires de séquences dans l'espace défini par 1024 composantes de l'ACP.

A partir du calcul de distance, les séquences sont séparées en deux classes en utilisant la classification hiérarchique basée sur la méthode de Ward. Cette méthode maximise l'inertie interclasse. La classification hiérarchique, faisant un usage important de la mémoire lorsque le jeu de données est trop grand, la classification est appliquée à 100 000 séquences choisies aléatoirement et une analyse discriminante linéaire (LDA) est utilisée pour classer les autres séquences.

2.2.3 Double-validation d'un sous-groupe

A chaque itération, la classification hiérarchique divise le jeu de données en deux groupes qui sont évalués avant d'être ou non redivisés. Afin d'éviter de diviser les données en familles trop petites, le premier critère de validation est la taille du sous-groupe. Si un groupe fait moins de 100 séquences, il n'est pas redivisé. Le deuxième critère de validation s'appuie sur le *matepair* et permet d'estimer la qualité de la partition. Ce terme correspond à la proportion de monomères ayant son plus proche voisin dans la même classe, se basant sur les distances euclidiennes calculées auparavant. Ainsi des valeurs de *matepairs* élevées (proches de 1) indiquent des sous-groupes bien homogènes et séparés validant la classification.

Le seuil de *matepair* est fixé à 0.90, pour avoir des groupes homogènes. Ce seuil est choisi d'après des études précédentes avec des monomères humains dont la classification était connue. Ce seuil est utilisé pour toutes les classifications, dans le but de faire une comparaison entre espèces. Si au moins une des valeurs de *matepair* est au-dessous de ce seuil, les sous-groupes sont considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.

2.3 Analyse des séquences

Les séquences monomériques sont comparées à partir de leur composition en 5-mers dans le but d'identifier des regroupements d' α -satellites sans passer par l'étape d'alignement. Pour chaque ensemble de monomères, la table de fréquence des 5-mers est analysée en utilisant l'Analyse en Composante Principale (ACP) pour réduire l'espace de complexité pour pourvoir visualiser les données sur les premiers plans factoriels.

L'alignement des séquences est fait avec Muscle [15] et visualisé avec Seaview [16]. La phylogénie est construite avec la méthode du maximum de vraisemblance (PhyML) [?]. Le modèle F84 est utilisé pour la construction de l'arbre. Le support de branche est aLRT (SH-like). La fréquence à l'équilibre des nucléotides, le ratio de transition et de transversion et les taux de variation sont optimisés.

Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B (TTCGTTGGAA[AG]CGGGA), PJ α (TTCCTTTT[CT]CACC[AG]TAG) et pK β (GATATCCGGTTCCCTT) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [17] et en autorisant 2 *mismatch* au maximum par rapport au consensus.

3 Résultats

3.1 Caractérisation intraspécifique des familles

Les séquences ont été classées en utilisant la méthode de classification objective développée dans l'équipe. Le nombre de familles est déterminé indirectement par un critère sélectionnant des classes à la fois homogènes et différentes les unes des autres.

3.1.1 Identification des familles

A l'issue de la classification, des familles de taille variable sont obtenues. Les familles sont fixées à un seuil de 100 séquences, cependant certaines familles "restantes" sont composées d'un tout petit nombre de séquences. Seules les familles supérieures au seuil de 100 séquences sont analysées. Pour les 4 espèces, le nombre de familles conservées diminue considérablement après élimination des petites familles (<100 séquences) mais ces familles ne représentent que 11% des séquences du jeu de données au plus (Table 1). Malgré le nombre de familles qui diffère d'une espèce à l'autre, la distribution des familles est similaire chez les quatre espèces. Les

TABLE 1 – Résumé du jeu de données et des résultats de la classification.

Espèces	Nombre d' α -satellites	Nombre de familles au total	Nombre de familles > 100 séquences	% de séquences dans les familles > 100 séquences
<i>C. solatus</i>	105 529	564	11	96.03
<i>C. pogonias</i>	112 902	132	13	98.71
<i>C. sabaeus</i>	29 842	338	43	89.11
<i>M. fascicularis</i>	235 535	3694	114	88.94

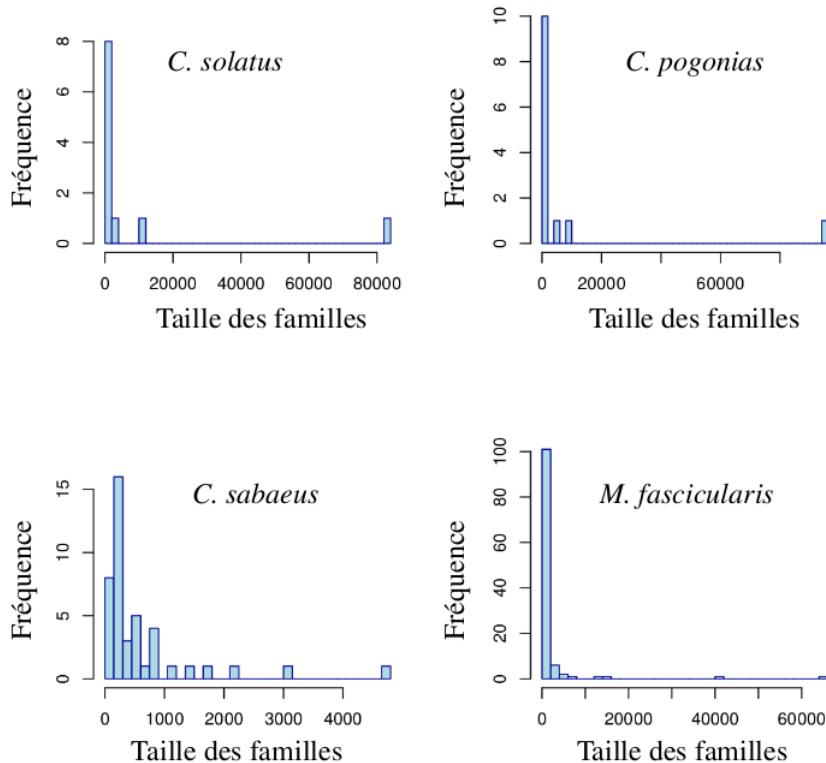


FIGURE 4 – Distribution des familles en fonction du nombre de séquences.

familles composées d'un grand nombre de séquences sont peu nombreuses par rapport aux familles composées de moins de 1 000 séquences (Figure 4).

Les espèces *C. solatus* et *C. pogonias* sont analysées dans un premier temps pour comparer la classification automatisée avec la classification empirique faite dans l'équipe. Six classes α -satellites ont été définies empiriquement et confirmées expérimentalement chez ces Cerco-pithèques. Ces deux espèces partagent deux grandes classes monomériques, nommées C1 et C2, et deux classes formant un HOR d'ordre 2, nommées C3-C4, de l'ordre d'une centaine de séquences chacune (données non montrées). *C. pogonias* possède les classes supplémentaires C5 et C6. Ces classes ont été définies manuellement à partir d'une ACP basée sur la composition en 5-mers de tous les monomères (Figure 5).

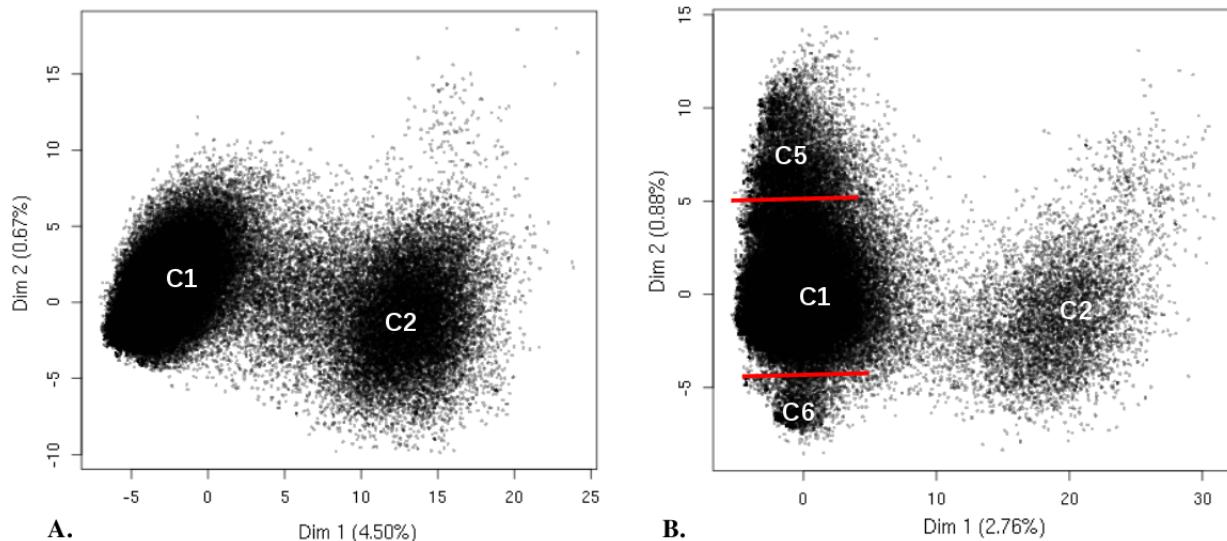


FIGURE 5 – Caractérisation visuelle des familles α -satellite chez *C. solatus* et *C. pogonias* à partir d'une ACP, basée sur la composition en 5-mers des monomères : Le nom des familles est indiqué sur les graphiques. Un point représente un monomère. **A.** *C. solatus*. **B.** *C. pogonias*.

TABLE 2 – Comparaison des classifications automatiques avec celles publiées précédemment [11], [12]. Les familles marquées d'une étoile sont des familles < 100 séquences, non conservées pour la suite des analyses.

Classes empiriques	Familles	
	<i>C. solatus</i>	<i>C. pogonias</i>
C1 (> 5 000 séquences)	1	9
C2 (> 82 000 séquences)	10	2
C3 (<170 séquences)	1*	1
C4 (<170 séquences)	1*	1*
C5 (8 300 séquences)	-	1
C6 (4 100 séquences)	-	0

Bien que le nombre de grandes familles soit relativement proche entre ces deux espèces (pour rappel 11 et 13), les résultats diffèrent significativement (Table 2). Chez *C. solatus*, les classes C1 à C4 sont retrouvées : 11 familles correspondent à C2, une famille correspond à C1 et les familles C3 et C4, ne comportant que 109 et 112 séquences dans le jeu de données, sont retrouvées dans des petites familles d'environ 80 séquences chacune. Chez *C. pogonias*, contrairement à C6, les classes C1, C2, C3, C4 et C5 sont retrouvées. C1 est répartie en 9 familles, alors que les classes C2, C3 et C5 sont regroupées dans des familles individuelles, et la classe C4 est également retrouvée sous la forme d'une petite famille de 86 séquences. La classe C2 chez le *C. solatus* et la classe C1 chez *C. pogonias* aurait dû former une famille chacune. Ces classes sont donc sur-clusterisées. Dans la suite de l'analyse, la pertinence de cette sur-clusterisation est examinée (plus particulièrement pour les séquences C2).

La diversité de ces classes a été analysée avec une ACP basée sur la composition en 5-mers des séquences (Figure 6) afin de valider la sur-clusterisation observée. Les familles, représentées par une couleur, sont superposées sur les classes empiriques (en noir, non visible).

Chez *C. solatus*, la classe C1 est entièrement retrouvée en une famille en vert. La classe C2 est répartie en 11 familles. Seules les 4 familles les plus grandes sont représentées par soucis de visibilité. Parmi celles-ci, deux familles intermédiaires (rouge et bleue) sont visibles entre la famille en vert et celle en orange. Elles se superposent. Une famille supplémentaire turquoise se démarque. Pour confirmer cette division des séquences de C2, la visualisation de l'ACP des 5-mers est observée en fonction des composantes 1 et 3. Les familles intermédiaires sont toujours confondues, contrairement à la famille turquoise qui forme visuellement une famille à part entière. Pour confirmer cette classification, 50 séquences aléatoires par famille (représentée sur l'ACP) sont utilisées pour construire un arbre phylogénétique, et chacune des familles est représentée par une couleur. Chaque couleur se regroupe, donc l'arbre atteste que chaque famille est bien retrouvée, notamment les familles intermédiaires (rouge et bleue) qui forment bien deux familles.

Chez *C. pogonias*, les classes C2, C3 et C5 sont bien retrouvées respectivement en orange, bleu foncé et jaune. La classe C6 n'est pas séparée de la classe C1 (en vert). La classe C1 est divisée en deux familles supplémentaires (en rose et violet). La visualisation des composantes 1 et 3 de l'ACP ne permet pas de trancher sur la l'individualité de ces deux familles. Pour cela, 50 séquences aléatoires par famille sont utilisées pour construire un arbre phylogénétique, et chacune des familles est représentée par une couleur. Chacune des couleurs se regroupe, justifiant

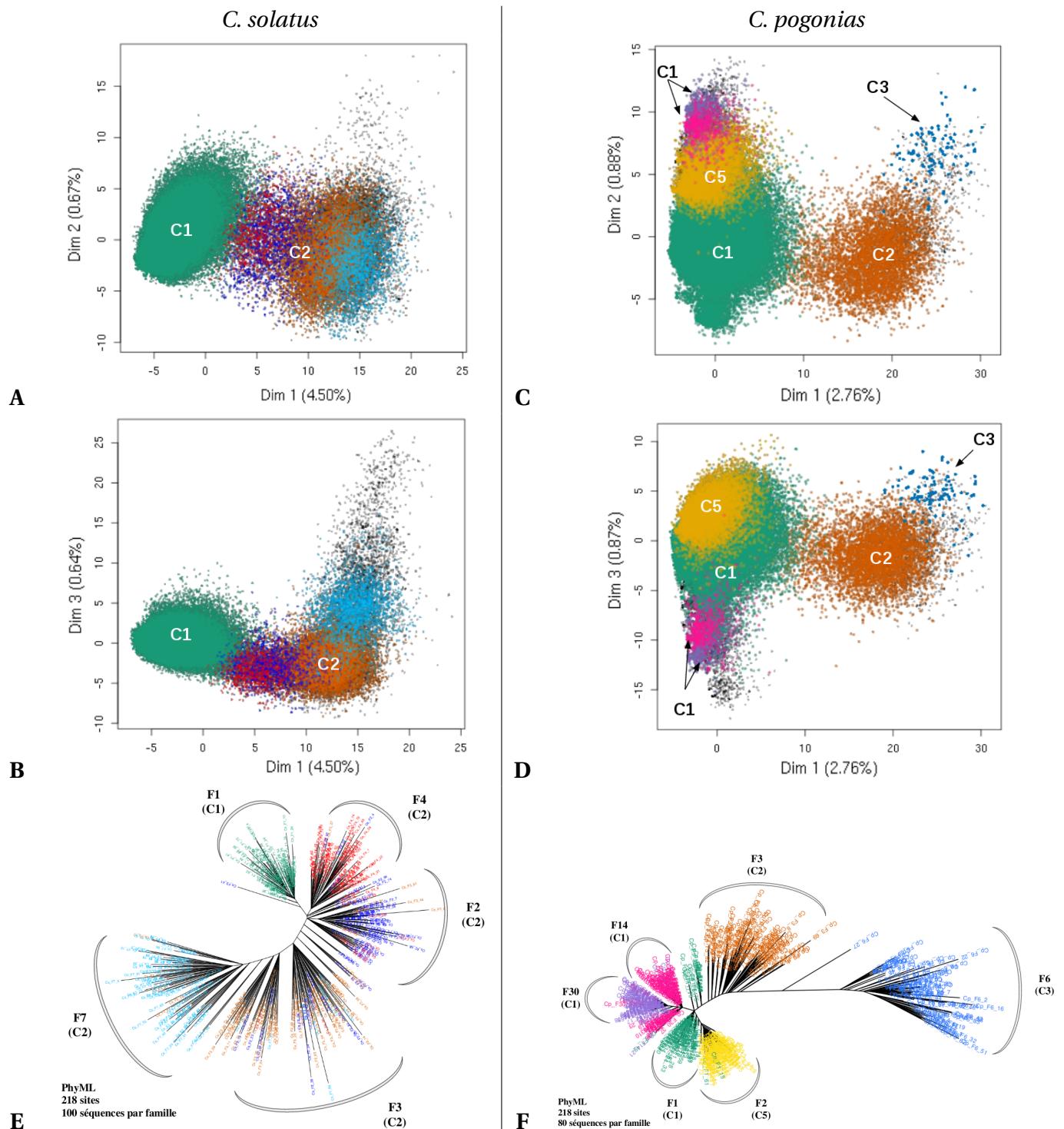


FIGURE 6 – Représentation des plus grandes familles issues de la classification automatisée :
 Ces familles sont superposées sur les représentations de l'ACP des 5-mers. **A.** Composantes 1 et 2 de l'ACP. Les familles qui correspondent à C1 sont en vert, C2 en orange, rouge, bleu et turquoise. **B.** Composantes 1 et 3 de l'ACP. **C.** Composantes 1 et 2 de l'ACP. Les familles qui correspondent à C1 sont en vert, violet et rose, C2 en orange, C4 en bleu clair et C5 en jaune. **C.** Composantes 1 et 3 de l'ACP. **E. et F.** Phylogénie des différentes familles chez *C. solatus* (100 séquences par famille) et *C. pogonias* (80 séquences par famille) respectivement. Les couleurs sont respectivement conservées.

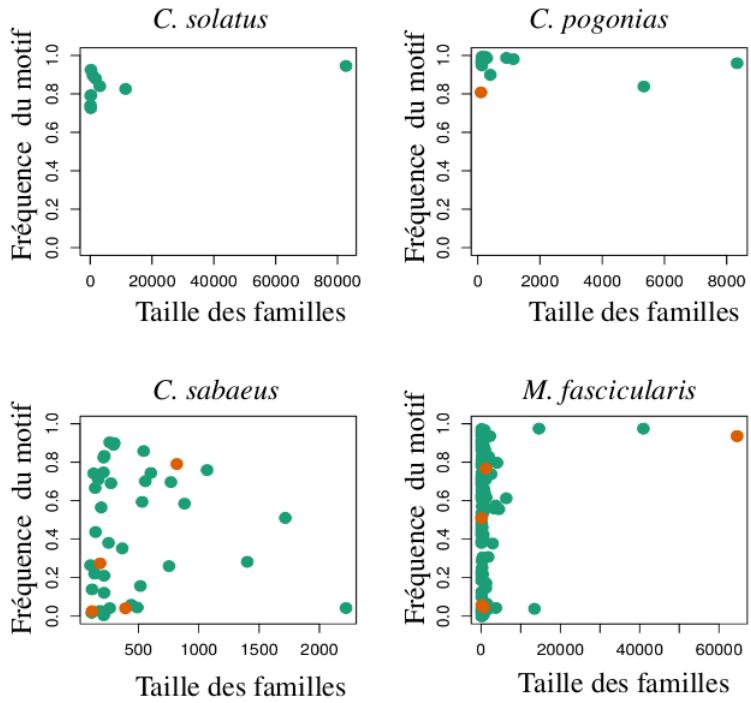


FIGURE 7 – Fréquences des motifs CENP-B, pJ α ou pK β dans les familles : Les motifs ont été cherchés sur la base de leur consensus en autorisant deux différences. Le pourcentage de séquences par famille ayant le motif pJ α est en vert, pK β en orange et CENP-B est absent de toutes les familles. Chaque famille est également représentée en fonction de sa taille.

la classification. Par ailleurs cet arbre montre que les familles en rose et violet sont très proches.

En conclusion, la classification automatique permet de retrouver les classes identifiées dans les travaux publiés, exceptée la classe C6 qui semble indissociable de la classe C1, et la mise en évidence de plusieurs familles dans les classes C2 chez *C. solatus* et C1 chez *C. pogonias*.

3.1.2 Motifs potentiellement fonctionnels

Ensuite je m'intéresse aux caractéristiques des familles α -satellites identifiées, notamment l'existence de motifs conservés. A ce jour, les motifs CENP-B et pJ α , capables de fixer respectivement les protéines CENP-B et pJ α (une protéine peu décrite), ont été décrits. Des travaux en cours dans l'équipe suggèrent l'existence d'un troisième motif conservé, nommé pK β , qui serait présent dans certaines familles. Ces trois motifs sont exclusifs l'un de l'autre. Ils sont situés au même endroit sur une séquence. Ils ont été recherchés pour chaque famille de chaque espèce sur la base de leur consensus, avec 2 *mismatch* autorisés (cf matériel et méthodes).

La protéine CENP-B est présente chez toutes les espèces de primates, mais les Cercopithèques ne possèdent pas son site de liaison. En effet, le motif CENP-B est absent chez *C. solatus* et *C. pogonias*. *C. sabaeus* et *M. fascicularis* n'ont pas ce motif non plus. Plus de 90% des familles

chez les quatre espèces ont le motif pJ α mais à des fréquences variables, et seulement 9% ont le motif pK β et quelques rares familles n'ont aucun des trois motifs chez *M. fascicularis* (Figure 7).

La totalité des familles chez *C. solatus* ont le motif pJ α à plus de 72%. Chez *C. pogonias*, 12 familles sur 13 ont le motif à plus de 83%. Une grande famille par espèce (> de 80 000 séquences), qui correspond à C1, a le motif à 95%. *C. sabaeus* présente 39 familles sur 43 avec ce motif, dont 7 l'ayant à plus de 75%. *M. fascicularis* a des pourcentages pour le motif pJ α qui varient entre 1% et 97%, dont 28 familles avec le motif à plus de 75%.

Une seule famille, incluant les séquences de C3, a ce motif à 80% chez *C. pogonias* et *C. solatus* (non représenté car famille < 100 séquences). *C. sabaeus* a quatre familles avec le motif, dont deux à 79% et l'autre à 29%. *M. fascicularis* a cinq familles avec ce motif. La famille ayant le motif à 93% est une grande famille de 64 000 séquences. Les deux autres familles ont le motif à 0.76% et 0.50%.

Pour conclure sur les α -satellites de ces espèces, le motif CENP-B est absent, le motif pJ α est le plus présent et le mieux conservé. Le motif pK β est, quant à lui, présent dans très peu de familles.

3.1.3 Similarité entre familles

Selon les hypothèses actuelles [18], les familles α -satellites seraient le résultat d'amplifications qui généreraient plusieurs centaines ou milliers de copies identiques au cours du temps. La divergence intra-famille a été estimée à partir d'un échantillon de 500 séquences aléatoires par famille afin d'estimer l'âge des familles, les familles anciennes ayant un taux de divergence élevé.

Lorsque les histogrammes sont observés dans la globalité, un seuil à 8% délimite le taux de divergence des familles. *C. solatus* et *C. pogonias* ont des pourcentages relativement faibles ne dépassant pas 15%, le *M. fascicularis* a des taux de divergence variant de 0.1% à 28%, et *C. sabaeus* a les valeurs les plus élevées allant de 11% à 28% (Figure 8).

Chez *C. solatus* et *C. pogonias*, les familles qui ont un taux de divergence inférieur à 8% correspondent à du C1 et C5 chez *C. pogonias*. Les familles qui ont un taux de divergence supérieur correspondent à du C2, et pour *C. pogonias* à C3 également. *C. sabaeus* n'a que des familles avec un taux de divergence élevé, supérieur à 10%. *M. fascicularis* a quelques familles avec un faible pourcentage de divergence et d'autres à 30%.

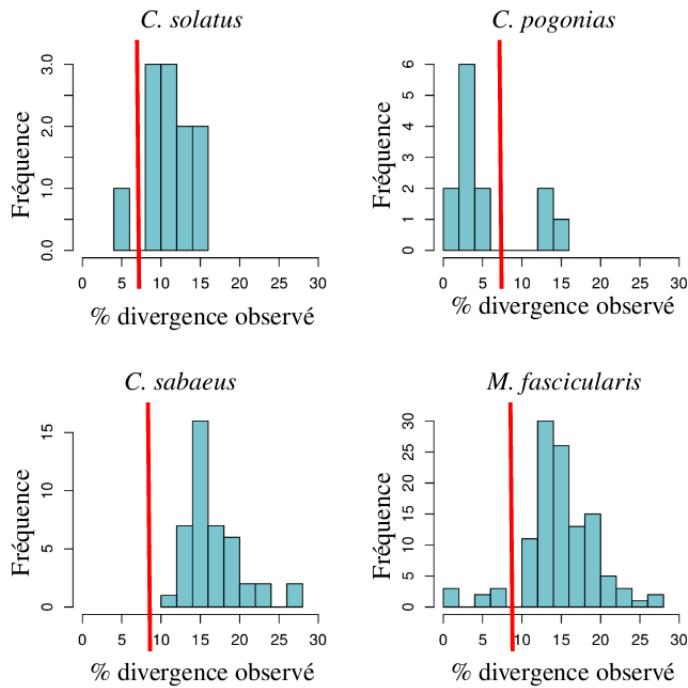


FIGURE 8 – Pourcentage de divergence observée dans les familles : Histogramme du nombre de familles en fonction du pourcentage de divergence observé au sein de celles-ci. Un même seuil est observé chez chaque espèces et il est indiqué en rouge sur chaque histogramme.

Pour conclure, une distribution bimodale est plus ou moins observée, avec des familles peu divergentes, notamment les familles qui correspondent à C1, et d'autres plus divergentes, notamment les familles qui correspondent à C2.

3.2 Comparaison inter-espèce

Pour comparer les classifications obtenues au sein de chaque espèce, une super-classification a été effectuée afin de déterminer l'existence de super-familles (SF). Pour cela, l'algorithme de classification a été utilisé sur un jeu de données composé de 100 séquences par grande famille (> 100 séquences) tirées aléatoirement pour chaque espèce. Un jeu de données de 18 100 séquences est ainsi créé à partir des 181 familles au total, puis soumis à la classification automatique. A l'issue de cette super-classification, 158 familles sont obtenues, dont 90 grandes familles (Figure 9). Les petites familles de moins de 20 séquences, soit 1.76% de ce jeu de données, ne sont pas prises en compte dans l'analyse. Après avoir présenté quelques caractéristiques de cette analyse, je présenterai les résultats concernant l'origine du site de fixation de pK β et la divergence des séquences appartenant à C2.

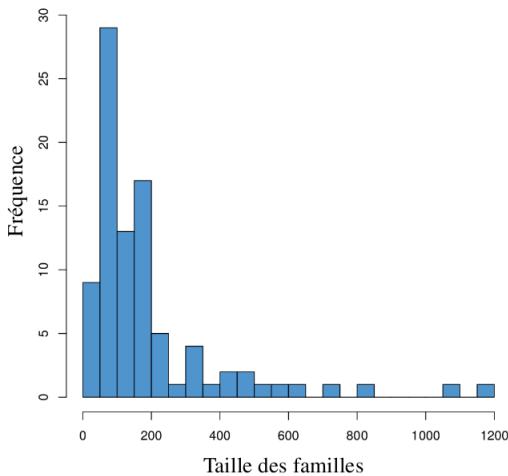


FIGURE 9 – **Distribution des super-familles :** Histogramme du nombre de familles observées en fonction de la taille.

3.2.1 Répartition des super-familles

Parmi les super-familles, seulement une *a priori* est commune aux quatre espèces. Elle rassemble 6 familles parmi les 10 familles classées C2 de *C. solatus* et les deux familles également classées C2 de *C. pogonias*, ainsi que deux familles du *M. fascicularis* et une famille du *C. sabaeus*. La classe C2, du moins une partie, serait donc commune aux quatre espèces.

Une super-famille est partagée entre *C. pogonias*, *C. sabaeus* et *M. fascicularis*. Cette super-famille regroupe des familles ayant le motif pK β uniquement et correspondrait à la classe C3 identifiée chez *C. pogonias* et *C. solatus*. Donc cette famille serait en fait commune aux 4 espèces également. Pour rappel, les séquences de la classe C3 de *C. solatus* n'ont pas été prises en compte à cause de la taille de la famille (< 100 séquences).

Certaines super-familles sont spécifiques aux espèces. Au sein des 8 super-familles spécifiques de *C. pogonias*, seule l'une d'entre elles regroupe deux familles, le reste étant composé d'une famille. Ces super-familles correspondent à 5 familles C1, une famille C5 et une petite partie de C6. Les trois super-familles spécifiques de *C. solatus* seraient équivalentes aux familles C2. *C. sabaeus* en aurait 3 et *M. fascicularis* 35.

C. sabaeus et *M. fascicularis* partagent 38 super-familles, soit 42% des grandes super-familles (> 20 séquences), dont la plus grande fait une taille de 1194 séquences. Une seule super-famille est uniquement commune à *C. solatus* et *C. pogonias* et elle regroupe les deux plus grandes familles qui correspondent à du C1.

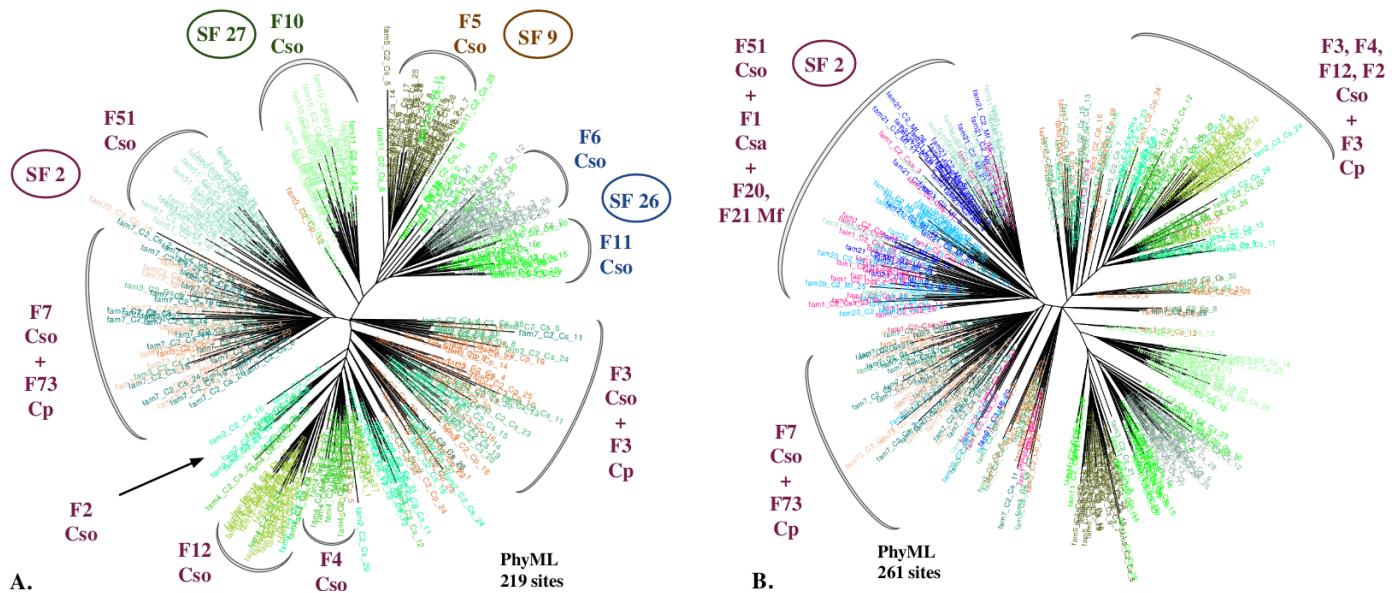


FIGURE 10 – Phylogénie des différentes familles qui correspondraient à la famille C2 : Les abréviations : F = famille, SF = super-famille, Cso = *C. solatus*, Cp = *C. pogonias*, Mf = *M. fascicularis*, Csa = *C. sabaeus*. Les super-familles sont écrites d'une même couleur. **A.** *C. solatus* (tons verts) et *C. pogonias* (tons orange) **B.** Des familles du *M. fascicularis* (tons bleus) et *C. sabaeus* (rose) sont rajoutées.

3.2.2 Un groupe C2 hétérogène

La classe C2 est divisée en 10 familles chez *C. solatus* et en 2 familles chez *C. pogonias*. Or dans les résultats empiriques publiés, les séquences C2 auraient dû former une seule famille chez chacune de ces deux espèces. La classe C2 est donc sur-clusterisée. Ensuite, certaines de ces familles se sont regroupées pour former une super-famille, notamment avec une famille de *C. sabaeus* et *M. fascicularis*.

Pour vérifier les résultats de cette super-classification, 30 séquences aléatoires de chacune des familles qui correspondent à C2 sont utilisées pour construire deux arbres phylogénétiques. Le premier contient uniquement les familles de *C. solatus* représentées en tons verts et celles de *C. pogonias* en tons orange (Figure 10 A). Le second contient les familles supplémentaires supposées de la classe C2 de *C. sabaeus* en rose et *M. fascicularis* en tons bleus (Figure 10 B).

Dans le premier arbre (Figure 10 A), chaque ton vert forme un groupe. Donc les familles de *C. solatus* forment bien des familles distinctes. Seule une famille semble se mélanger aux autres. Deux de ces familles contiennent des tons orange. Donc *C. solatus* et *C. pogonias* partagent deux familles. Cette sur-clusterisation aurait donc du sens. Par contre, la super-famille regroupe trop de familles sans faire de distinction. Dans le second arbre (Figure 10 B), les séquences du *M.*

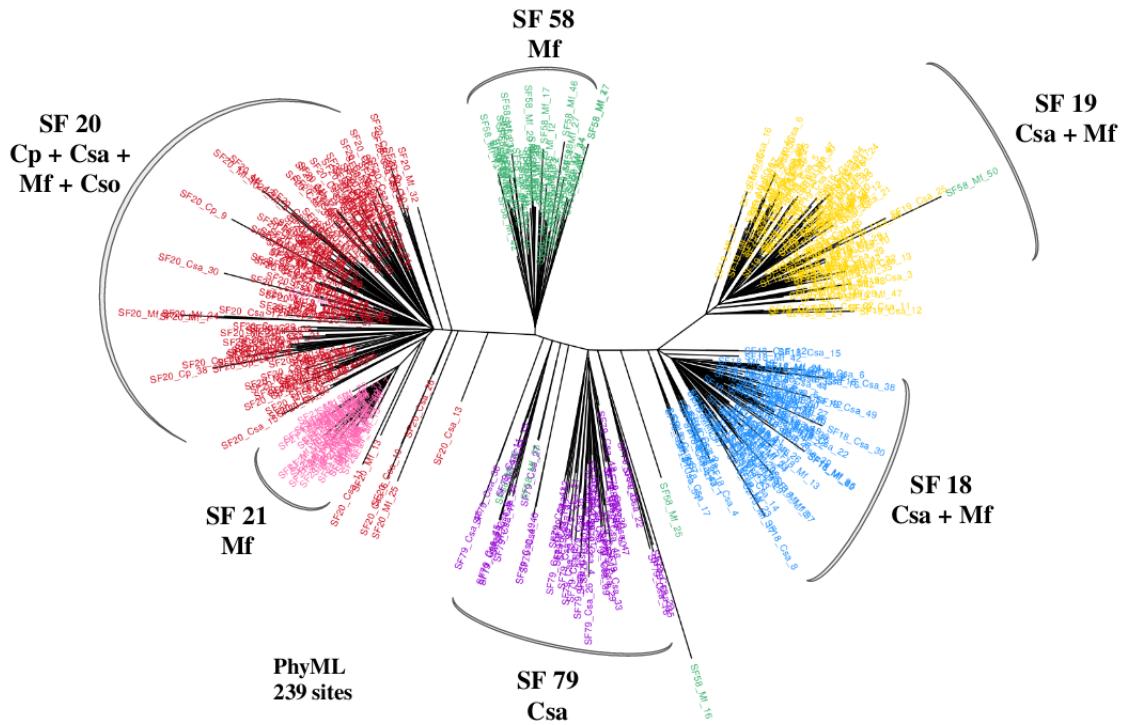


FIGURE 11 – Phylogénie des super-familles pK β : Les abréviations : SF = super-famille, Cp = *C. pogonias*, Cso = *C. solatus*, Csa = *C. sabaeus*, Mf = *M. fascicularis*. Une couleur correspond à une super-famille.

fascicularis et *C. sabaeus* vont spécifiquement se mettre avec une famille de *C. solatus*.

Pour conclure sur l'hétérogénéité de C2, une partie de ces familles est spécifique à *C. solatus*, deux sont communes à *C. solatus* et *C. pogonias*, une famille de *C. solatus* est commune aux quatre espèces selon la classification. Cependant la super-classification ne sépare pas aussi bien les familles.

3.2.3 Origine du motif pK β

Pour poursuivre l'étude sur les familles communes entre espèces, le regroupement des familles α -satellites ayant le motif pK β ou "familles pK β " attirent l'attention.

La première super-famille pK β regroupe une famille de chacune des espèces. Ces familles pK β possèdent le motif à 80%. De plus cette famille correspond à la classe C3. Donc toutes ces espèces auraient une "famille C3" en commun. Le *M. fascicularis* et *C.sabeus* ont deux super-familles pK β en commun. Les familles constituant ces super-familles ont un pourcentage de présence du motif relativement faible, pourtant elles sont tout de même rassemblées. *C.sabeus* et *M. fascicularis* ont respectivement une et deux super-familles pK β spécifiques. La première super-famille a le motif à 27% et la deuxième à 50%.

Un arbre composé de toutes les familles pK β est construit pour voir si les super-familles sont retrouvées phylogénétiquement. Le jeu de données est construit à partir de 50 séquences aléatoires par famille pK β par espèce (Figure 11). Une couleur représente une super-famille. Chaque couleur se regroupe, donc les super-familles sont retrouvées. Les séquences se rassemblent exactement comme dans les super-familles.

Pour conclure, les familles pK β sont des familles partagées entre espèces pour certaines, et spécifique à une espèce pour d'autres. La méthode de classification retrouve bien les familles pK β (intra-spécifique) et les super-familles pK β inter-spécifique).

4 Discussion

Après classification, les séquences α -satellites sont réparties en familles. Le nombre de familles par espèce est très variable. *C. solatus* et *C. pogonias* ont seulement un dizaine de familles comparé au *M. fascicularis* qui en a un peu plus d'une centaine, et *C. sabaeus* a un nombre intermédiaire. Il semble donc, qu'au sein des primates, il y ait une réelle hétérogénéité en terme de nombre de familles. Il se pourrait aussi, chez *C. solatus* et *C. pogonias* que certaines familles n'aient pas été digérées par les deux enzymes de restriction utilisées, et par conséquent que le nombre de famille chez ces deux espèces soit sous-estimé.

La méthode de classification crée beaucoup de petites familles (< 100 séquences). Plusieurs séquences sont alors éliminées et pourraient correspondre à une "vraie" famille. Ces familles de petite taille sont probablement des séquences ayant subi des réarrangements, ayant une délétion importante, ou dont la qualité est moins bonne. Néanmoins, leur poids dans l'analyse des familles α -satellites serait négligeable.

Parmi les familles analysées, la méthode de classification a retrouvé la majorité des classes empiriques, mais elle n'a pas réussi à dissocier les séquences C6 de C1. Ce groupe n'est pas retrouvé même si la valeur seuil du paramètre *pairmate* est augmenté ou diminué. Par contre, elle a sur-clusterisé C1 chez *C. pogonias* et C2 chez *C. solatus*. Cet algorithme de classification aurait donc du mal à séparer des séquences qui seraient trop proches mais il pourrait séparer un groupe qui n'est pas parfaitement homogène. En effet, les divisions de la classe C2 ont bien été retrouvées sur un arbre phylogénétique pour *C. solatus*. Peut-être qu'une notion de sous-groupes devrait être définie pour affiner l'analyse.

Chez ces familles, le motif le mieux conservé est pJ α . Il est présent dans pratiquement toutes

les familles, même si sa présence est moindre dans quelques unes des familles chez *C. sabaeus* et *M. fascicularis*. Le motif de pK β , bien que largement moins présent, reste très conservé. La conservation du motif n'a rien à voir avec la taille de la famille. Par exemple, pour le motif pK β , une grande famille de plus de 60 000 séquences possède le motif à 93% chez *M. fascicularis* et une famille de 104 séquences chez *C. pogonias* a le motif à 80%.

Pour estimer l'âge de ces séquences, le taux de divergence au sein de chaque famille a été calculé. Chez *C. solatus* et *C. pogonias*, les séquences C1 ont le pourcentage de divergence le plus faible, elles sont donc des familles récentes situées au cœur du centromère, d'après le gradient d'âge. Les séquences de la famille C2 et C3, chez *C. pogonias*, ont les pourcentages de divergence les plus élevés. Ce sont les familles les plus anciennes, qui ont accumulé des mutations, et qui seraient localisées au niveau du péri-centromère. Une distribution bimodale est observée, qui pourrait supposer des vagues d'amplification. De plus, la super-classification regroupe des familles des quatre espèces qui correspondent à du C2. Le taux de divergence, autour de 14 %, s'accordent avec cette information. La méthode de classification confirme l'analyse phylogénétique.

C. sabaeus avait seulement des familles avec un fort taux de divergence. Ces résultats seraient liés à l'assemblage. En effet, on n'aurait que des séquences péricentromériques. A l'inverse, les données du *M. fascicularis* contiennent non seulement des reads courts mais également des séquences de l'assemblage. L'un correspondrait à des séquences provenant du cœur du centromère, et l'autre du péri-centromère.

Certaines familles de différentes espèces auraient le même taux de divergence. La question se poserait de savoir, par exemple, si les familles qui correspondent à C1 chez *C. solatus*, *C. pogonias* et *M. fascicularis* seraient communes aux quatre espèces. D'après des études récentes au sein de l'équipe, il semblerait que le Macaque n'ait pas de familles de la classe C1. Mais peut-être que cette famille serait présente chez *C. sabaeus*. Puisque les familles récentes ne sont pas disponibles dans le jeu de données et que cette espèce est très proche des deux autres Cercopithèques, cette hypothèse serait à vérifier.

La méthode de classification, confirmée par l'analyse phylogénétique, rassemble une famille de *C. sabaeus* et deux familles du *M. fascicularis* avec des séquences C2 de *C. solatus* et *C. pogonias*. Ces espèces ont donc des séquences qui seraient potentiellement des séquences C2. Les taux de divergence relativement élevés, autour de 14% pour ces familles, s'accordent avec cette information. Cependant, il pourrait y avoir d'autres séquences de C2 qui ne seraient pas

identifiées.

L'apparition des familles pK β pourrait être déduite avec le pourcentage de divergence au sein de ces familles. Parmi celles-ci, les plus anciennes sont la super-famille spécifique à *C. sabaeus* et celle spécifique au *M. fascicularis*, avec un pourcentage de divergence de 17,50% et 16,12% respectivement. La famille la plus jeune est la deuxième super-famille spécifique au macaque avec 7,09% de divergence. La super-famille commune aux 4 espèces est relativement ancienne, avec en moyenne 14,70% de divergence, et les familles communes à la fois à *M. fascicularis* et au *C. sabaeus* sont un peu plus jeunes avec respectivement 12,50% et 11,80% de divergence. Pourtant les familles les plus anciennes auraient dû être retrouvées chez chacune des espèces. Soit ces familles n'ont pas été détectées chez toutes les espèces, soit ces familles ont disparu au cours du temps.

Pour évaluer la reproductibilité, plusieurs super-classifications ont été lancées et analysées. Avec les mêmes paramètres, d'une super-classification à l'autre, les super-familles obtenues sont différentes, notamment pour la classe C2. En effet, les familles C2 vont se regrouper différemment, formant parfois plus ou moins de super-familles. Cette répartition dépend certainement de la qualité des séquences sélectionnées aléatoirement. Seules les familles pK β se rassemblent exactement de la même manière. Cette méthode de classification possède donc un problème de reproductibilité pour certaines familles, notamment la C2.

5 Conclusion

Pour conclure, les séquences α -satellites ont été réparties en familles, identifiées en fonction de ce qui a été trouvé dans la littérature. La présence des motifs fonctionnels pJ α et pK β sont les mieux conservées dans ces familles. L'âge des familles a été estimée avec pourcentage de divergence au sein des familles. Certaines familles sont récentes et d'autres anciennes mais l'espèce *C. sabaeus* manquerait d'information dans le cœur du centromère. Une fois les familles caractérisées, la super-classification a permis de trouver les familles communes entre espèces et de déduire les familles les plus anciennes et les plus récentes ou encore celles qui sont spécifiques aux espèces. *M. fascicularis* et *C. sabaeus* partageraient le plus de familles. Deux familles communes aux quatre espèces ont été définies : les super-familles pK β et "C2". La méthode de classification met en avant l'existence de familles, confirmées phylogénétiquement, mais certaines familles trop proches ne sont pas détectées, et d'autres sont regroupées différemment

pendant la super-classification.

Il faudrait améliorer l'algorithme pour essayer d'affiner la détection de familles ou encore pour améliorer la reproductibilité des familles. Il serait aussi intéressant d'étudier la structure en HOR ou en organisation monomérique. Si les familles sont définies en une structure particulière, la question se pose de savoir si elles conservent la même structure d'une espèce à l'autre.

Références

- [1] Don W Cleveland, Yinghui Mao, and Kevin F Sullivan. Centromeres and kinetochores : from epigenetics to mitotic checkpoint signaling. *Cell*, 112(4) :407–421, 2003.
- [2] Stefano Santaguida and Andrea Musacchio. The life and miracles of kinetochores. *The EMBO journal*, 28(17) :2511–2531, 2009.
- [3] Kevin F Sullivan, Mirko Hechenberger, and Khaled Masri. Human cenp-a contains a histone h3 related histone fold domain that is required for targeting to the centromere. *The Journal of cell biology*, 127(3) :581–592, 1994.
- [4] S Henikoff, K Ahmad, and H S Malik. The centromere paradox : stable inheritance with rapidly evolving dna. *Science (New York, N.Y.)*, 293 :1098–1102, August 2001.
- [5] Giuliana Giannuzzi, Claudia Rita Catacchio, and Mario Ventura. Centromere evolution : Digging into mammalian primary constriction. In *Current Frontiers and Perspectives in Cell Biology*. InTech, 2012.
- [6] Valery A Shepelev, Alexander A Alexandrov, Yuri B Yurov, and Ivan A Alexandrov. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS genetics*, 5(9) :e1000641, 2009.
- [7] H F Willard. Evolution of alpha satellite. *Current opinion in genetics & development*, 1 :509–514, December 1991.
- [8] D M Kurnit and J J Maio. Variable satellite dna's in the african green monkey cercopithecus aethiops. *Chromosoma*, 45 :387–400, May 1974.
- [9] I Alexandrov, A Kazakov, I Tumeneva, V Shepelev, and Y Yurov. Alpha-satellite dna of primates : old and new families. *Chromosoma*, 110 :253–266, August 2001.
- [10] CR Catacchio, R Ragone, G Chiatante, and M Ventura. Organization and evolution of gorilla centromeric dna from old strategies to new approaches. *Scientific reports*, 5 :14189, 2015.
- [11] Lauriane Cacheux, Loïc Ponger, Michèle Gerbault-Seureau, Florence Anne Richard, and Christophe Escudé. Diversity and distribution of alpha satellite dna in the genome of an old world monkey : Cercopithecus solatus. *BMC genomics*, 17 :916, November 2016.
- [12] Lauriane Cacheux, Loïc Ponger, Michèle Gerbault-Seureau, François Loll, Delphine Gey, Florence Anne Richard, and Christophe Escudé. The targeted sequencing of alpha satellite

dna in cercopithecus pogonias provides new insight into the diversity and dynamics of centromeric repeats in old world monkeys. *Genome biology and evolution*, June 2018.

- [13] Mark S Springer, Robert W Meredith, John Gatesy, Christopher A Emerling, Jong Park, Daniel L Rabosky, Tanja Stadler, Cynthia Steiner, Oliver A Ryder, Jan E Janečka, Colleen A Fisher, and William J Murphy. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PloS one*, 7 :e49521, 2012.
- [14] Andrej Benjak, Claudia Sala, and Ruben C Hartkoorn. Whole-genome sequencing for comparative genomics and de novo genome assembly. *Methods in molecular biology (Clifton, N.J.)*, 1285 :1–16, 2015.
- [15] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput, 2004.
- [16] N Galtier, M Gouy, and C Gautier. Seaview and phylo_win : two graphic tools for sequence alignment and molecular phylogeny. *Computer applications in the biosciences : CABIOS*, 12 :543–548, December 1996.
- [17] Peter Rice, Ian Longden, and Alan Bleasby. Emboss : the european molecular biology open software suite, 2000.
- [18] M Katharine Rudd, Gregory A Wray, and Huntington F Willard. The evolutionary dynamics of alpha-satellite. *Genome research*, 16 :88–96, January 2006.

Résumé

Chez les primates, des séquences répétées en tandem dans le centromère encore appelées ADN α -satellite. Un monomère α -satellite fait 171 pb de long. Ces séquences ont un taux d'identité de 60% à 100%. Ces monomères peuvent être regroupés en familles selon la similarité. Pour étudier ces familles, des méthodes classiques, comme la phylogénie, existent mais elles ne permettent pas de classer un grand nombre de séquences. De plus les méthodes ne sont pas objectives et ne permettent pas de comparer les espèces entre elles. Pour pallier ce problème, l'équipe ARChE a développé une méthode de classification permettant de traiter des centaines de milliers de séquences. Mon sujet consiste à appliquer cette méthode aux jeux de données déjà publiés pour évaluer la méthode. Dans un deuxième temps, cette méthode est appliquée à deux autres primates dans le but de caractériser les familles d'espèces proches. Les mécanismes d'évolution sont déduits à partir d'une comparaison inter-espèce révélant les différences et les familles communes.

Abstract

Centromeric repeated sequences in Primates are named α -satellite DNA. An α -satellite's length is about 171 pb. Its identity rate is between 60% and 100%. Monomers can be gathered into families according to their similarity. To study these families, methods using phylogeny are used but a large number of sequences cannot be processed. Moreover these methods are not objective and do not allow inter-species comparison. To overcome this problem, the team ARChE developed a classification method which processes hundred of thousands of sequences. The subject of my internship is to apply this method to published datasets and assess this method. Then this method is applied to other Primates in order to characterize families in close species. Mechanism of evolution are deducted from inter-species comparison, revealing common and different families.

Annexes

